

MENU

SEARCH

INDEX

DETAIL

NEXT

1/2



JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11)Publication number: 10187666

(43)Date of publication of application: 21.07.1998

(51)Int.Cl.

G06F 17/18

G06F 17/30

(21)Application number: 08349487

(71)Applicant:

NEC CORP

(22)Date of filing: 27.12.1996

(72)Inventor:

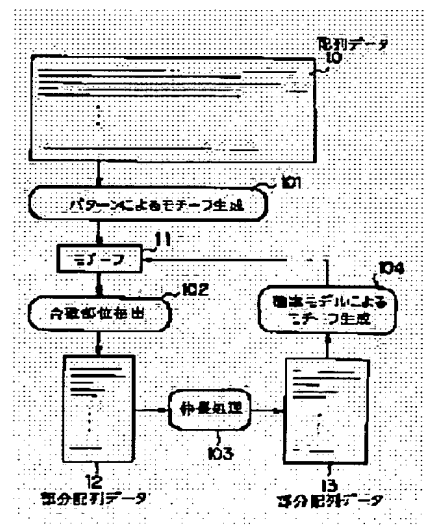
KUROIWA YUKIKO

(54) METHOD FOR FINDING PROBABILISTIC MOTIF OF ARRAY OF PROTEIN AND GENE

(57)Abstract:

PROBLEM TO BE SOLVED: To find a motif which can represent repetition positively with a probabilistic model in a short time when the motif is searched for from amino acid array data on protein and base array data on a gene.

SOLUTION: A motif is generated with a pattern from given array data 10 (step 101), a parts matching the motif 11 is extracted as partial array data 12 from the array data 10 (step 102), and parts before and after the partial array data 12 are extended to generate new partial array data 13 (step 103). Then a motif 11 is generated with probability on the basis of the partial array data 13 by using a hidden Markov model (step 104). The steps 102 to 104 are repeated to find a probabilistic motif.



LEGAL STATUS

[Date of request for examination] 27.12.1996

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number] 2907169

[Date of registration] 02.04.1999

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998 Japanese Patent Office

[MENU](#)

[SEARCH](#)

[INDEX](#)

[DETAIL](#)

[NEXT](#)

(51)IntCl.⁵

識別記号

F I

G 0 6 F 17/18
17/30G 0 6 F 15/36
15/40Z
3 7 0 F

審査請求 有 請求項の数7 O L (全 9 頁)

(21)出願番号 特願平8-349487

(22)出願日 平成8年(1996)12月27日

(71)出願人 000004237

日本電気株式会社

東京都港区芝五丁目7番1号

(72)発明者 黒岩 由希子

東京都港区芝五丁目7番1号 日本電気株
式会社内

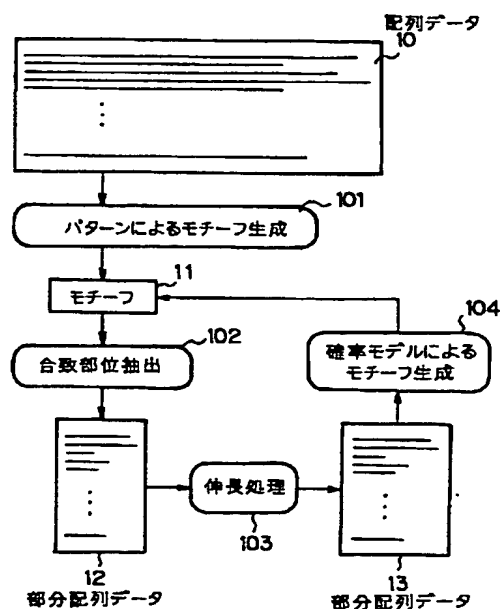
(74)代理人 弁理士 若林 忠

(54)【発明の名称】 タンパク質及び遺伝子の配列の確率的モチーフ発見方法

(57)【要約】

【課題】タンパク質に対するアミノ酸配列データや遺伝子に対する塩基配列データからモチーフを探索する際に、確率的モデルにより、反復を陽に表現することができるモチーフを短時間で発見できるようにする。

【解決手段】与えられた配列データ10からパターンによってモチーフを生成し(ステップ101)、配列データ10の中からモチーフ11に合致する部位を抽出して部分配列データ12とし(ステップ102)、部分配列データ12の前後を伸長して新たに部分配列データ13を生成する(ステップ103)。その後、部分配列データ13に基づき、隠れマルコフモデルを用いて確率的にモチーフ11を生成する(ステップ104)。ステップ102～ステップ104を繰り返すことにより、確率的モチーフを発見する。



【特許請求の範囲】

【請求項1】 タンパク質及び遺伝子のいずれか一方の配列データから確率的モチーフを発見する方法において、

与えられた配列データから記号処理によって抽出したモチーフを初期のモチーフとして、前記配列データの中から前記初期のモチーフに最も適合する部分配列を第1の部分配列として抽出する第1のステップと、
前記第1の部分配列の前後を延長して新たに第2の部分配列を生成し、前記第2の部分配列を確率的に表現したものを新たにモチーフとする第2のステップと、
前記配列データの中から前記第2のステップで得たモチーフに最も適合する部分配列を抽出して改めて前記第1の部分配列とする第3のステップとを有し、
前記第1のステップの実行後、前記第2のステップ及び前記第3のステップを繰り返し実行することにより前記確率的モチーフを発見することを特徴とする、確率的モチーフ発見方法。

【請求項2】 前記第1のステップにおいて、アミノ酸及び塩基のいずれか一方の並びであるパターンに関し前記配列データでの出現数を計算し、その出現数が多い方から選択したパターンを前記初期のモチーフとする、請求項1に記載の確率的モチーフ発見方法。

【請求項3】 前記第2のステップにおいて前記モチーフを隠れマルコフモデルを用いて表現し、前記第3のステップにおいて前記配列データから前記モチーフに最も適合する部分配列を取り出す際には、配列長で尤度を正規化し、正規化後の尤度が高い部分配列を取り出す、請求項1に記載の確率的モチーフ発見方法。

【請求項4】 前記第2のステップにおいて、前記配列データを用いて前記モチーフを確率的に生成する際に、所与の状態数からなる隠れマルコフモデルを初期モデルとし、遷移確率が所定値以下の遷移を取り除き新たな状態を結合して新たな隠れマルコフモデルのネットワーク構造を生成し、

その後、生成した前記ネットワーク構造の隠れマルコフモデルの初期パラメータをランダムに複数の組設定することと、前記設定に対してパラメータ算出を行うことと、前記パラメータ算出で最良の結果を与える隠れマルコフモデルを新たな隠れマルコフモデルとすることとを繰り返

し、
確率的にモチーフを生成する請求項3に記載の確率的モチーフ発見方法。

【請求項5】 前記第2のステップにおいて、隠れマルコフモデルの状態数及びネットワーク構造を最適化するために、前記配列データを学習用データと評価用データとに分け、前記配列データのうち前記学習用データのみを隠れマルコフモデルの学習の際に使用し、学習後の隠れマルコフモデルの中から前記評価用データの尤度が最も高い状態数及びネットワーク構造を探索し、探索され

た状態数及びネットワーク構造に基づいて前記パラメータを再推定し、確率的にモチーフを生成する、請求項4に記載の確率的モチーフ発見方法。

【請求項6】 前記第2のステップにおいて、尤度の低い部分配列を予め定めた割合で除去してから前記モチーフを生成する、請求項3に記載の確率的モチーフ発見方法。

【請求項7】 隠れマルコフモデルの状態数及びネットワーク構造を情報量基準で最適化する請求項3に記載の確率的モチーフ発見方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、タンパク質及び遺伝子の配列におけるモチーフを探索する方法に関し、特に、配列データから確率的モチーフを発見する方法に関する。

【0002】

【従来の技術】モチーフとは、タンパク質のアミノ酸配列や遺伝子の塩基配列において、構造上・機能上の重要な部位であり、生物の進化の過程において保存されている部分配列のことである。モチーフはコンセンサス配列とも呼ばれている。

【0003】従来、タンパク質のアミノ酸配列のモチーフ及び遺伝子の塩基配列のモチーフは、パターンと呼ばれる表現で表わされていた。例えば、タンパク質モチーフのデータベースであるPROSITEでは、“N-(P)-[ST]-P”や“[ST]-x-[ST]”などのように、20種のアミノ酸のOR（論理和）の並びとしてモチーフを表現していた。ここで、N、P、S、Tは、アミノ酸の種類を表わす略号であって、それぞれ、アミノ酸の一種であるアスパラギン酸、プロリン、セリン、トレオニンを示している。また、xは任意のアミノ酸を示し、{P}は、P（プロリン）以外のアミノ酸全てを示し、[ST]は、S（セリン）またはT（トレオニン）を示す。

【0004】また、一部では、プロファイルと呼ばれる表現方法も用いられている。これは、モチーフの各位置でそれぞれのアミノ酸または塩基が来た時のスコアを表現したものである。さらに一方では、タンパク質や遺伝子の配列を隠れマルコフモデル（HMM：Hidden Markov Model）のような確率的モデルによって表現する手法が使用されつつある。

【0005】このような表現を用いて、生化学的に関係のある異なる配列から自動的に共通部位を発見できれば、この共通部位は生化学的な機能や立体構造のために重要な部分配列であって進化的に保存されてきたものであると考えることができる。このため、タンパク質や遺伝子の機能や構造の解析のために、共通部位であるモチーフの発見は重要な問題である。モチーフを発見する手法としては、これまでにも、ボトムアップによる手法、

トップダウンによる手法など、さまざまな手法が研究されている。

【0006】ボトムアップによる手法は、パターンの空間を考えて、それぞれのパターンに対して配列での出現頻度または情報量 (Information Content) を計算し、出現頻度または情報量の多いパターンをモチーフとする手法である。例えば、1990年発行の米国の雑誌 "Proc. Natl. Acad. Sci. USA" の第87号826-830頁記載の Smithらの論文 "Finding sequence motifs in groups of functionally related proteins" や、1989年発行の米国の同雑誌の第86号1183-1187頁記載の Stormoらの論文 "Identifying protein-binding sites from unaligned DNA fragments" などがある。

【0007】トップダウンによる手法は、配列を比較してローカルな類似度を見つめる手法であり、配列のアライメントを行なう。アライメントとは、与えられた複数の配列に、ギャップなどを入れて、同じあるいは性質の似たアミノ酸または塩基が縦のカラム位置に並ぶようにする手法である。トップダウンによる手法には、配列を2本ずつペアワイズにアライメントを行ない、その結果を組み合わせる手法や、グループ間でアライメントする手法がある。このようなアライメントを用いたモチーフ抽出では、配列をアライメントしてから同じアミノ酸または塩基が揃っている部位を探すか、類似の局所的な部位をアライメントする。このようなアライメントとしては、例えば、Blast, FastA, Clust V などがある。

【0008】

【発明が解決しようとする課題】 しかしながら、従来のボトムアップによる手法では、計算時間がかかることが問題点であった。例えば、文字数 $| \Sigma |$ での長さ l のパターンでは、探索空間のサイズは、 $O(| \Sigma |^l)$ となり、パターンの長さ l で指数関数的に大きくなる。さらに、実際にタンパク質モチーフのデータベースである PROSITE を見てみると、進化的に保存されている部位であるとは言え、配列におけるギャップやミスマッチは少なくない。ボトムアップによる手法においてギャップやミスマッチを考慮しようとすると、ギャップやミスマッチを許すような近隣語も考えなければならなくなり、探索空間が大きくなり過ぎて、計算時間の観点から実現が難しい。しかも、得られるパターンの表現能力は、隠れマルコフモデルなどの確率的モデルよりも低い。

【0009】 また、従来のアライメントによる手法では、初期の段階の誤りが波及して精度があまり良くなかったり、配列の重み付けを考えなければならなかったり、位置に依存したスコアを考えづらかったりという問題点があった。

【0010】 本発明は、上記問題点の解決を図り、共通

な部位だけに注目してギャップやミスマッチを確率的モデルを用いて表現することで、配列を詳細に表現できる確率的モチーフの発見手法を提供することを目的とする。さらに本発明は、モチーフを隠れマルコフモデルで表現することで、反復するモチーフを陽に扱うことのできる確率的モチーフの発見手法を提供することを目的とする。

【0011】

【課題を解決するための手段】 本発明の確率的モチーフ発見方法は、請求項1に記載のように、タンパク質及び遺伝子のいずれか一方の配列データから確率的モチーフを発見する方法において、与えられた配列データから記号処理によって抽出したモチーフを初期のモチーフとして、配列データの中から前記初期のモチーフに最も適合する部分配列を第1の部分配列として抽出する第1のステップと、第1の部分配列の前後を延長して新たに第2の部分配列を生成し、第2の部分配列を確率的に表現したものを新たにモチーフとする第2のステップと、配列データの中から第2のステップで得たモチーフに最も適合する部分配列を抽出して改めて第1の部分配列とする第3のステップとを有し、第1のステップの実行後、第2のステップ及び第3のステップを繰り返し実行することにより確率的モチーフを発見する。

【0012】 請求項2に記載の発明は、請求項1に記載の確率的モチーフ発見方法において、第1のステップにおいて、アミノ酸及び塩基のいずれか一方の並びであるパターンに関し配列データでの出現数を計算し、その出現数が多い方から選択したパターンを初期のモチーフとする。また、請求項3に記載の発明は、請求項1に記載の確率的モチーフ発見方法において、第2のステップにおいてモチーフを隠れマルコフモデルを用いて表現し、第3のステップにおいて配列データからモチーフに最も適合する部分配列を取り出す際には、配列長で尤度を正規化し、正規化後の尤度が高い部分配列を取り出す。

【0013】 さらに請求項3に記載の確率的モチーフ発見方法では、第2のステップにおいて確率的にモチーフを生成する際に、①所与の状態数からなる隠れマルコフモデルを初期モデルとし、遷移確率が所定値以下の遷移を取り除き新たな状態を結合して新たな隠れマルコフモデルのネットワーク構造を生成し、その後、生成したネットワーク構造の隠れマルコフモデルの初期パラメータをランダムに複数の組設定することと、設定に対してパラメータ算出を行うことと、パラメータ算出で最良の結果を与える隠れマルコフモデルを新たな隠れマルコフモデルとすることとを繰り返したり、②隠れマルコフモデルの状態数及びネットワーク構造を最適化するために、配列データを学習用データと評価用データとに分け、配列データのうち学習用データのみを隠れマルコフモデルの学習の際に使用し、学習後の隠れマルコフモデルの中から評価用データの尤度が最も高い状態数及びネットワーク構

造を探索し、探索された状態数及びネットワーク構造に基づいてパラメタを再推定したり、③尤度の低い部分配列を予め定めた割合で除去してからモチーフを生成したりすることができる。さらに請求項3に記載の確率的モチーフ発見方法では、隠れマルコフモデルの状態数及びネットワーク構造を情報量基準で最適化することができる。

【0014】（作用）本発明では、パターンよりも表現能力の高い確率的モデルを用いることによって、モチーフ内のギャップやミスマッチの扱いを容易にし、モチーフを精密に表現することができる。特に、隠れマルコフモデルを使って表現することによって、反復するモチーフも陽に扱えるようにした。確率モデルとして隠れマルコフモデルを用いてモチーフを生成する際には、例えば、特開平7-36920号公報に記載のタンパク質配列データに対する隠れマルコフモデル作成方法を用いて、隠れマルコフモデルのトポロジー及びパラメタを学習する。

【0015】モチーフ発見の最初から確率的モデルを適用するのでは、対象とするタンパク質やアミノ酸の配列データのデータ量が一般に大きいと、計算時間がかかり過ぎる。アミノ酸配列の配列長、すなわち対象とするアミノ酸の残基数は、通常、300~400程度だが、最大で5000を超える配列も存在する。また、遺伝子の配列は、使用される文字数は少ないものの、一般にはアミノ酸配列よりさらに長い。そこで、本発明では、計算時間を短縮するため、最初のステップでは、記号処理によって、例えばパターンで大まかにモチーフの部分に絞り、その後、それをもとに確率的モデルによる計算を実行する。そして、確率モデルによるモチーフ生成を何度か繰り返すことで、次第に配列の伸長を行ないながら、モチーフを詳細に表現する。

【0016】

【発明の実施の形態】次に、本発明の実施の形態について、図面を参照して説明する。図1は、本発明によるタンパク質及び遺伝子の配列の確率的モチーフ発見方法の実施の一形態を説明するための図である。なお、各図において、通常の矩形（角が直角となっている矩形）はデータを表わし、角が丸くなっている矩形は処理を表わし、菱形は処理における判断と分岐を表わしている。

【0017】本実施の形態では、モチーフ発見の対象となる配列データとして、タンパク質におけるアミノ酸配列あるいは遺伝子における塩基配列の配列データ10が与えられる。配列データ10には、通常、複数の配列に関するデータが含まれる。まず、この配列データ10に対して、記号処理の1種であるパターン処理によるモチーフ生成を実行し、モチーフ11を抽出する（ステップ101）。次に、合致部位抽出処理を実行して（ステップ102）、配列データ10中でのモチーフ11に合致する部位を抽出し、抽出された部位を部分配列データ

12とする。伸長処理を実行し（ステップ103）、この部分配列データ12の前後を伸長して部分配列データ13とする。そして、伸長後の部分配列データ13に対し、確率モデルによるモチーフ生成処理を実行し（ステップ104）、確率モデルで表現されたモチーフ11を生成してステップ102に戻る。本実施の形態では、ステップ102~104の過程を繰り返すことによって、確率的モチーフを発見する。

【0018】以下、各処理について、図面を参照しながらさらに詳しく説明する。

【0019】図2は、ステップ101におけるパターンによるモチーフ生成処理に使用する表の一例を示している。この表50には、パターンと、そのパターンが出現した配列数（出現数）を書き込む。同一の配列に同一のパターンが2度以上出現した場合にも、出現数は1と数えることとする。これは、一部の配列に特異的に出現する共通部位を除くためである。したがって、表中の出現数の値は、常に、与えられた配列データ10に含まれる配列数以下となる。ここに例示した表は、ロイシンジッパーモチーフを含む159個の配列からモチーフ発見を行なった時に得られたパターンとその出現数の一部を表わしている。

【0020】本実施の形態において、パターンは、定まったアミノ酸もしくは塩基と、任意のアミノ酸もしくは塩基との並びの形式とする。すなわち、[ST]や

{P}などのアミノ酸もしくは塩基の論理和（OR）は認めないものとする。これは、計算時間の都合上、論理和や情報量は考慮せず、最もシンプルなパターンだけを考えるということである。このようにシンプルなパターンを用いた抽出でも、本実施の形態では後で確率的モデルで詳細に表現するので、構わない。

【0021】パターンにおいて定まっているアミノ酸もしくは塩基の数と、任意のアミノ酸もしくは塩基xの長さの最大は、予め適当に決めておく。例えば、アミノ酸もしくは塩基の数を3、xの長さの最大を3に決めた場合のパターンの例は、“LxxExxxL”や“TxxxTA”などである。ここで、L,Eは、それぞれ、アミノ酸の一種であるロイシン、グルタミン酸を示し、T,Aは、それぞれ、塩基の一種であるチミン、アデニンを示す。

【0022】図3は、ステップ101でのパターンによるモチーフ生成処理において、図2に示すような表を用いるべく多くの配列に含まれるパターンを選ぶための処理を示すフローチャートである。配列データ20（＝配列データ10）を上から順に見て、パターン生成処理（ステップ111）によりパターンを一つずつ生成し、そのパターンが同一配列中で新規かどうかを判断する（ステップ112）。ステップ112において、生成されたパターンが新規でない場合には、ステップ116に移行し、生成されたパターンが新規の場合には、そのパ

ターンが既に表(図2参照)に記載されたパターンであるかどうか、すなわち、そのパターンが表内で新規かどうかを判断する(ステップ113)。表内で新規でない場合には、表でのそのパターンの出現数に1を加算(インクリメント)し(ステップ114)、ステップ116に移行する。表内で新規の場合には、その表にそのパターンを書き込むとともにその出現数を1として(ステップ115)、ステップ116に移行する。アミノ酸配列や塩基配列では、配列中に全てのパターンが現れるわけではないので、本実施の形態では、新規のパターンが現れる度に、そのパターンを表に書き込み、表を拡張することとし、これによって、表を保持するためのメモリを節約する。

【0023】ステップ116では、配列データ20中の配列を全部取り出したか、すなわちデータ終了かどうかを判断し、データ終了でなければ上述の処理を繰り返すためにステップ111に戻って次のパターン生成に移り、データの終了となったら、表での出現数が上位のパターン21を選出する(ステップ117)。このパターン21は、モチーフ11として、合致部位の抽出など次の処理で使用される。

【0024】図4は、モチーフ11がパターンで表現されていた場合の合致部位探索のための処理(ステップ102)を示すフローチャートである。ここでは、モチーフ11がパターン31として与えられるものとする。なお、モチーフ11が確率モデルで表わされている場合の合致部位探索処理については、図7及び図8を用いて後述する。パターンが含まれない配列からは部分配列を抽出しないままとし、多くのパターンが含まれている配列からは多くの部分配列を抽出することとする。また、同一部位が同時に複数のパターンに合致している場合には、その部位を複数回抽出するのではなく、一つの部分配列として抽出することとする。

【0025】与えられた配列データ30(=配列データ10)に対して、部分配列生成処理を実行して、順次、部分配列32を生成する(ステップ121)。そして、この部分配列32が、与えられたパターン31のいずれかに合致するかを比較する(ステップ122)。いずれかに合致する場合には、その部分配列32をファイルに書き込んで(ステップ123)、ステップ124に移行し、合致しない場合にはそのままステップ124に移行する。ステップ124では、データ終了かどうかを判断し、データ終了でなければ、ここで述べた過程を繰り返すためにステップ121に戻る。ステップ124でデータ終了であれば、合致部位探索の処理を終了するが、この時点で、パターン31に合致する部分配列33がファイルに得られている。

【0026】次に、伸長処理(ステップ103)について説明する。伸長処理は、配列データ10から探索・抽出された部分配列データ12に対し、その部分配列の前

後にある長さで配列要素(塩基やアミノ酸)を付加して、部分配列を伸長する処理である。実際には、合致部位抽出の際に、併せて伸長処理を行なえばよい。この伸長処理は、現在のモチーフの付近にある進化的な保存部位を探すためだけでなく、パターンの開始位置・終了位置を揃えるために行なっている。例として、前述したロイシンジッパーモチーフを発見する場合を考える。このモチーフは、周期7でL(ロイシン)が現れる“(Lxxxxxx)ⁿL”という形式である。パターンによるモチーフ生成で、アミノ酸の数を3、xの長さの最大を3に決めた場合、前述した通り図2の表に示す結果が得られていた。ここで伸長処理を行わないものとする

と、“KxxExL”や“LxKxxE”で抽出した部分配列は、“LxxxxxxL”のパターンで抽出した部分配列と異なるデータとして学習されるが、伸長処理を行うことによって、それらのパターンで抽出した部分配列を揃えて学習することができる。

【0027】次に、確率モデルによるモチーフ生成処理(ステップ104)について説明する。ここでは、確率モデルとして隠れマルコフモデル(HMM:Hidden MarkovModel)を使用する場合を説明する。以下、隠れマルコフモデルのことを、HMMと略記する。

【0028】ところで、特開平7-36920号公報には、タンパク質配列データに対するHMM作成方法が開示されている。同公報での実施例に示されているように、タンパク質配列データに対するHMM作成方法は、一般の信号にも応用が可能である。そこで本実施の形態では、特開平7-36920号公報に記載の方法をタンパク質及び遺伝子の配列に適用できるように改良して、モチーフを表現したHMMを部分配列から作成するようにした。図5は、本実施の形態でのHMMの作成過程を表わすフローチャートである。なお、作成過程の細部については、特開平7-36920号公報を参照のこと。

【0029】まず、HMMの状態数の最大 N_{max} を予め与え、与えられた部分配列のデータ(伸長処理後の部分配列データ13)を学習データと評価データとに分ける(ステップ131)。また、初期HMMのネットワーク構造を、少ない状態数の全結合プラス終了状態のHMMとする(ステップ132)。次に、予め試行回数 $CNTMAX$ が指定されているとして、初期パラメタを乱数とし、学習データに対してBaum-Welchアルゴリズムによるパラメタ推定を試行回数 $CNTMAX$ だけ繰り返す。iに現在のHMMの状態数、候補Mに尤度が最大のHMMを代入する(ステップ133)。ここで、学習データ Y_1, Y_2, \dots, Y_M に対する尤度とは、 $Prob(Y_i)$ をHMMが Y_i を出力する確率としたとき、

【0030】

【数1】 $Prob(Y_1) \times Prob(Y_2) \times \dots \times Prob(Y_M)$

のことである。そして、現在のHMMの状態数iが、予

め指定した最大状態数 N_{max} に達したかどうかを判断する(ステップ134)。 $i \geq N_{max}$ の場合は、ステップ137に移行して後処理を行う。一方、 i が N_{max} になっていない場合には、現在の状態数 i でのHMMである $bstM_i$ に上述の候補Mを保存し(ステップ135)、候補Mから確率が ϵ 未満の遷移を除き状態を付加したHMMを初期HMMのネットワーク構造とし(ステップ136)、ステップ133に戻り、上記の過程を繰り返す。ここで ϵ は、予め定めたいきい値である。

【0031】ステップ137の後処理では、それぞれの状態数 i のHMMである $bstM_i$ に対して評価データでの尤度を計算し、評価データでの尤度が最大となる状態数のHMMを選び、計算に使われない状態を消去する(ステップ138)。すなわち、初期確率がゼロより大の状態からいくつかの遷移を通っても到達できない状態、及び、いくつかの遷移を通っても最終状態に到達できない状態をそれぞれ除く。こうして得られたHMMの状態数とネットワーク構造のデータ35により、学習データと評価データの両方を使ってパラメタを再推定し(ステップ139)、確率モデルによるモチーフを生成する。

【0032】以上、確率モデルによるモチーフ生成処理の一例を説明したが、確率モデルによるモチーフ生成処理は上述した処理に限定されるわけではない。図6は、確率モデルによるモチーフ生成の別の例を示すフローチャートである。以下、図6に示すモチーフ生成処理について説明する。

【0033】まず、HMMの状態数の最大 N_{max} を予め与え(ステップ141)、初期HMMのネットワーク構造を、少ない状態数の全結合プラス終了状態のHMMとする(ステップ142)。次に、予め試行回数 $CNTMAX$ が指定されているとして、初期パラメタを乱数とし、部分配列データに対してBaum-Welchアルゴリズムによるパラメタ推定を試行回数 $CNTMAX$ だけ繰り返し、 i に現在のHMMの状態数、候補Mに尤度が最大のHMMを代入する(ステップ143)。そして、現在のHMMの状態数 i が、予め指定した最大状態数 N_{max} に達したかどうかを判断する(ステップ144)。 $i \geq N_{max}$ の場合は、ステップ147に移行して後処理を行う。一方、 i が N_{max} になっていない場合には、現在の状態数 i でのHMMである $bstM_i$ に上述の候補Mを保存し、 MDL_i にMの記述長を保存し(ステップ145)、候補Mから確率が ϵ 未満の遷移を除き状態を付加したHMMを初期HMMのネットワーク構造とし(ステップ146)、ステップ143に戻り、上記の過程を繰り返す。

【0034】ステップ147の後処理では、 MDL_i が最大となる M_i を選び(ステップ148)、上述した例と同様の計算によって、 M_i から使われない状態を消去し、使われない状態が消去された M_i を確率モデルによ

るモチーフとする。 MDL_i は候補 M_i の記述長を表わしているから、結局、情報量基準によって、HMMの状態数及びネットワーク構造が最適化されたことになる。

【0035】以上、HMMによる確率モデルによりモチーフを作成する処理を説明した。HMMによれば、ギャップやミスマッチを容易に表現することができる。ギャップはHMMの自己ループなどで表現でき、ミスマッチは各状態の出力確率で表現できる。さらに、本実施の形態あるいは特開平7-36920号公報に記載のHMM作成方法では、left-to-right型に限らず一般のネットワーク構造を学習できるため、反復するモチーフなども陽に扱うことができる。

【0036】ところで、実際にモチーフ発見を行なう場合には、初めに与えられる配列データに、モチーフを全く含まない配列が含まれている場合もある。これには、誤ってノイズとしてモチーフが与えられる場合も考えられるし、与えられたグループのサブグループのモチーフを抽出しようとしている場合も考えられる。これらの場合に対応するために、尤度 $Prob(Y_i)$ が低くなる部分配列 Y_i を予め定めておいた割合で除くといよい。例えば、図5で示した例の場合、ステップ137で評価データの尤度を計算する際に、尤度の低い部分配列を予め定めておいた割合で除いて尤度を計算する。また、このステップ137において、選んだ状態のネットワーク構造のHMMのパラメタを学習データと評価データの両方を用いて再推定する際にも、尤度の低い部分配列は除いてパラメタ推定する。図6で示した例の場合も同様に、後処理(ステップ147)において、尤度の低い部分配列は除いてから再びパラメタ推定をするようにしてもよい。

【0037】次に、合致部位探索処理(ステップ102)のうち、モチーフ11が確率モデルで表わされている場合の処理を説明する。図7は、モチーフが確率モデルで表現されていた場合に合致部位抽出のために使う表51の一例を示している。

【0038】表51では、それぞれの配列に対して、その配列に含まれる部分配列のなかで配列長ごとに尤度が最大となる配列が、その最大尤度とともに書き込まれる。実際には、この合致部位探索処理に引き続く伸長処理のために、部分配列を伸長し伸長後の部分配列を表51に書き込むようにしてもよい。配列長の最大と最小は予め与えておく。例えば、タンパク質のモチーフの場合、配列長の最小は3、最大は50程度で十分である。図示した例では、最大尤度だけを書き込んでいるが、配列にモチーフが複数出現する状態に対応するため、配列長ごとに、尤度が上位である部分配列とその尤度を書き込む表を用いてもよい。

【0039】図8は、モチーフが確率モデルで表現されていた場合における、合致部位抽出のための処理を説明するフローチャートである。ここでは、処理の前半で候

補を絞ることと正規化のための前処理を行ない、後半で尤度を正規化して候補から部分配列を選んでいく。

【0040】配列長（モチーフ長）の最大と最小を示すデータ41が予め与えられているとして、まず、配列データ40（＝配列データ10）を順に見て、配列長の最大と最小（データ41）にしたがって、順次、部分配列43を生成する（ステップ151）。次に、この部分配列43に対する、与えられた確率モデル42の尤度（または確率）44を計算する（ステップ152）。例えば、確率モデル42としてHMMを用いている場合には、フォワードバックワード（Forward-Backward）アルゴリズムにより、部分配列の尤度を計算すればよい。

【0041】次に、配列長で尤度44を正規化するための前処理を実行する（ステップ153）。この前処理と平行して、尤度44と、図7に示した表において対応する配列番号、配列長での尤度とを比較し（ステップ154）、得られた尤度44の方が表に記載された最大尤度よりも大きくない場合にはステップ156に移行し、尤度44が最大尤度より大きい場合には、表において対応する部分配列と尤度とを書き換えてから（ステップ155）、ステップ156に移行する。ステップ156では、配列データ40中の配列を全部取り出したか、すなわちデータ終了かどうかを判断し、データ終了でなければ上述の処理を繰り返すためにステップ151に戻って次の部分配列生成に移る。一方、ステップ156でデータの終了となったら、その時点で、各配列と各配列長に対して最大尤度になる部分配列とその尤度とが表に得られているから、ステップ153での前処理の結果を用い、正規化のための計算を行う（ステップ157）。正規化処理では、前処理によって得られた定数を使用して表の尤度を正規化する。そして、各配列に対して正規化した尤度が最大となる部分配列を取り出す（ステップ158）。このようにして得られた部分配列データ45が、合致部位探索の結果として得られた部分配列データである。

【0042】さて、正規化のための定数A、Bは、尤度の全データからも求めることができるが、データ量が膨大なために、計算時間や計算のための記憶領域を過度に必要とする。そこで、ステップ153での前処理では、得られた尤度のデータ44をまとめておく。例えば、ランダムに選択されたデータのみを保存して、ステップ157ではこの保存されたデータから定数A、Bが算出されるようにする。あるいは、各部分配列の尤度を配列長で分けて、さらに適当な数ごとにグループにして、それらのグループの尤度の平均、尤度の標準偏差を保存する。これらのまとめたデータを用いて、ステップ157では、配列長により尤度を正規化する。

【0043】配列長により尤度を正規化した値は、Z Scoreと呼ばれる。このZ Scoreは、以下のようにして求める。今、部分配列Yの尤度Lを正規化す

ることを考える。この部分配列Yの対数尤度 $\log L$ と配列長SeqLenは、定数A、Bを用いて、

【0044】

$$\text{【数2】 } \log L = A \times \text{SeqLen} + B$$

と線形近似することができる。ステップ153（前処理）で処理しながら集めた尤度のデータから、ステップ157での正規化のための定数A、Bを求め、次式により尤度を正規化した値 $\log L(N)$ を求める。

【0045】

10 【数3】

$$\log L(N) = \frac{\log \text{Prob}(Y)}{A \times \text{SeqLen} + B}$$

この $\log L(N)$ の平均 $AV(N)$ と標準偏差 $SD(N)$ を求め、部分配列YのZ Scoreを

【0046】

【数4】

$$Z\text{Score} = \frac{\log L(N) - AV(N)}{SD(N)}$$

とする。Z Scoreの平均は0.0、標準偏差は1.0である。

【0047】本実施の形態では、図1においてステップ102～104で表わした過程を繰り返すことによって、確率的モチーフを発見する。合致部位抽出（ステップ102）において前の繰り返しの時とほとんど同じ部分配列を抽出した場合、抽出した部分配列の長さがあらかじめ指定した配列長を越えた場合、あるいは、あらかじめ指定した回数以上の繰り返しがあった場合に、この繰り返しを終了する。

【0048】

30 【発明の効果】以上説明したように本発明は、確率的モチーフを発見するため、パターンによる表現よりも精密にモチーフを表現することができる。また、最初に、パターンなどの記号処理によって配列データを大まかにモチーフの部分に絞り、その後、隠れマルコフモデルなどによる学習を実行するので、確率的モチーフの発見方法でありながら、モチーフ発見に時間がかかりすぎるのを防ぐことができる。

【図面の簡単な説明】

40 【図1】本発明によるタンパク質及び遺伝子の配列の確率的モチーフ発見方法の実施の一形態を説明するフローチャートである。

【図2】パターンによるモチーフ生成処理に使用する表の一例を示す図である。

【図3】パターンによるモチーフ生成処理を説明するフローチャートである。

【図4】モチーフがパターンで表現されていた場合の合致部位探索の処理を説明するフローチャートである。

【図5】確率モデルによるモチーフ生成の過程の一例を示すフローチャートである。

50 【図6】確率モデルによるモチーフ生成の過程の別の例

13

14

を示すフローチャートである。

【図7】モチーフが確率モデルで表現されていた場合に合致部位抽出のために使用する表の一例を示す図である。

【図8】モチーフが確率モデルで表現されていた場合の合致部位抽出の処理を説明するフローチャートである。

【符号の説明】

10, 20, 30, 40 配列データ

11 モチーフ

12, 13, 45 部分配列データ

21, 31 パターン

32, 33, 43 部分配列

35, 41 データ

42 確率モデル

44 尤度

50, 51 表

101~104, 111~117, 121~124 ス

テップ

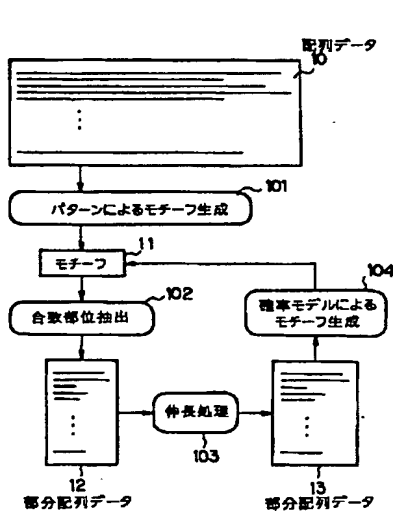
131~139, 141~148, 151~158 ス

10 テップ

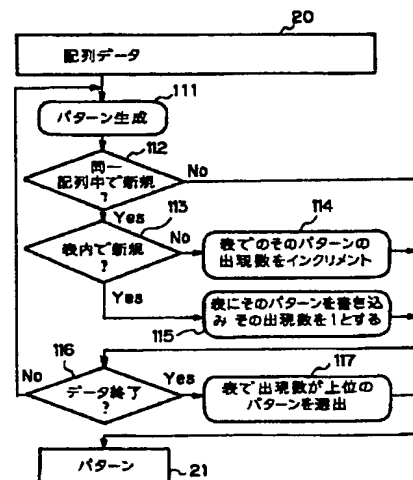
【図1】

【図2】

【図3】

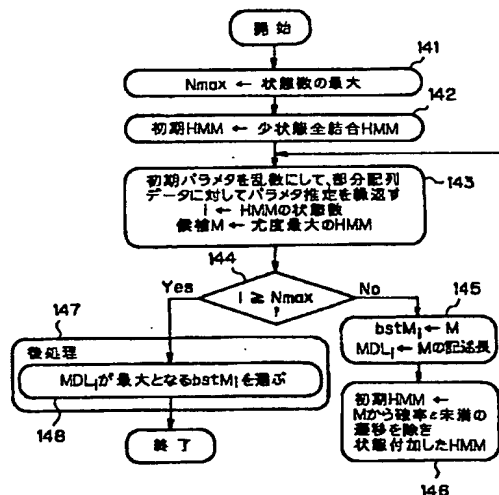
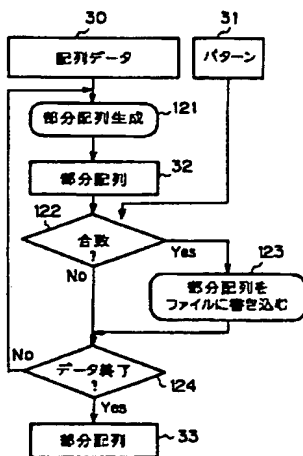


パターン	出現数
LxxExxxL	108
LxxxbxxL	89
LxxQxxL	85
KxxExL	84
LxxKxxE	83
:	:

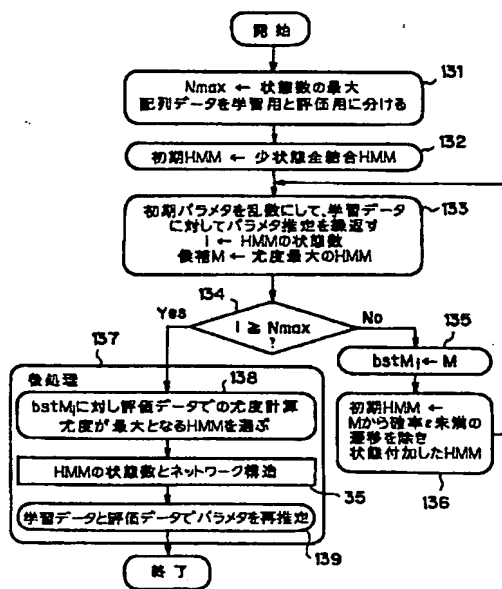


【図4】

【図6】



【図5】



【図7】

51

配列番号	配列長	部分配列	尤度
1	3	<div style="width: 25%;"></div>	<div style="width: 25%;"></div>
1	4	<div style="width: 33%;"></div>	<div style="width: 33%;"></div>
1	5	<div style="width: 40%;"></div>	<div style="width: 40%;"></div>
:	:	:	:
1	50	<div style="width: 100%;"></div>	<div style="width: 100%;"></div>
2	3	<div style="width: 25%;"></div>	<div style="width: 25%;"></div>
:	:	:	:
m	50	<div style="width: 100%;"></div>	<div style="width: 100%;"></div>

【図8】

